DOCUMENT RESUME

ED 299 322 TM 012 357

AUTHOR Coffman, William E.
TITLE Measurement of Thinking Skills--An Historical Perspective. Iowa Testing Programs Occasional Papers. Number 34.
PUB DATE Jun 88
NOTE 21p.; Paper presented at a Meeting to discuss future directions for the California Assessment Program (Berkeley, CA, July 16-18, 1987).
PUB TYPE Speeches/Conference Papers (150) -- Information Analyses (070)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cognitive Measurement; Cognitive Processes; Cognitive Tests; *Critical Thinking; Elementary Secondary Education; Higher Education; *History; Literature Reviews; *Measurement Techniques; Problem Solving; *Psychometrics; Test Construction
IDENTIFIERS *California Assessment Program; Higher Order Skills; *Thinking Skills

ABSTRACT

A historical review of educational literature concerning measurement of cognitive skills is provided, with reference to test construction. Focus is on assessing abilities beyond recall of information. In light of interest in assessing critical thinking in the California Assessment Program, it must be noted that concern with critical thinking is not a new development. As early as 1937, C. H. Judd anticipated the need for the kinds of cognitive studies that currently are of interest. California has been in the forefront of the movement toward an emphasis on thinking skills as goals of education. Some approaches abandoned in the past might be feasible as California plans the new assessment program. Promising ideas may be found in the work of R. Tyler (1930, 1932) and some of his associates. Another source might be the work of the Board of Examiners at the University of Chicago (Illinois), beginning in 1937. A similar source is the work of the Eight Year Study of the Progressive Education Association (1942) or the cooperative study of Evaluation in General Education (1954). The work of E. F. Lindquist might be examined to yield approaches to critical thinking assessment. Tests from the Educational Testing Service and the National Assessment of Educational Progress might also be useful sources. A 60-item reference list is provided. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED299322

IOWA TESTING PROGRAMS OCCASIONAL PAPERS

Number 34 - June 1988

Measurement of Thinking Skills -- An Historical Perspective

William E. Coffman

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

William E. Coffman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper prepared for a meeting in Berkeley CA on July 16-18, 1987 to discuss future directions for the California Assessment Program

TM012357

Measurement of Thinking Skills -- An Historical Perspective

William E. Coffman

Even a cursory look at recent books and issues of journals in education reveals a major emphasis on the teaching and measurement of thinking skills (Anderson, 1981; Brandt, 1984, 1985; Paul, 1984; Sternberg, 1985a, 1985b; Tyler & White, 1979; Whimsey, 1984). However, if one looks back at listings for earlier years, one finds about the same numbers under "critical thinking" as in current years. Concern with critical thinking on the part of educators is not a new development in the 1980's. Rather, what is new is concern with the possibility that recent research in cognitive psychology may be providing a foundation necessary for developing more successful ways of teaching thinking skills than has been the case in years past.

As early as 1937, Judd anticipated the need for the kinds of cognitive studies that have energized current interest. He wrote:

> Statistical studies have served to bring into educational psychology many illuminating facts regarding the common characteristics of large numbers of individuals, but only slow and laborious laboratory investigations and case studies can achieve the penetrating analyses of the mental activities of individuals (p. 146).

He clearly believed that thinking skills could be taught and blasted away at the advocates of a simple content emphasis in the schools. In Judd's words:

> When one looks at the writing of advocates of the new formalism for hopeful signs of interest in the higher mental processes, one is disappointed to find nothing but repeated references to heredity as the sole source of genius and to reflex arcs in the nervous system as the only physical basis of art appreciation or scientific thinking (p. 149).

And he pointed out the sorts of things he felt schools should be doing.

> ....the major task of the schools is to attempt to teach all individuals to make comparisons and draw contrasts, to look for explanations which will bring together ideas, to apply the systematic modes of thinking which the race has evolved, and to express clearly in language the relations between events and between facts. The schools have not adequately performed this task (p. 193).

The issue of whether it is possible to teach thinking skills directly that will generalize across different fields of study, however, is far from resolved. So far, the burden of proof is on the shoulders of those who would develop special courses in problem solving, critical thinking, or the like. Certainly, there seems to be general agreement that an individual has to think about some kinds of content, and without knowledge of the content, sound thinking is unlikely. And since most teachers, particularly those in secondary schools or colleges, have been trained in one or a limited number of traditional subject fields, instruction is likely to be concerned with the content of those fields--at least for the time being.

Even without the developments in cognitive psychology, it is possible that the current revival of emphasis on thinking skills might still have taken place. A review of the educational literature of the 20th century reveals an ebb and flow of concern on the part of educators for attention to critical thinking, or higher mental processes, or understanding as opposed to information or knowledge or memory. This movement points up a basic recurring issue in educational philosophy, i.e., should the schools be concerned primarily with the transmission of knowledge, or should they

be devoting attention to developing skills in using knowledge for social purposes (Coffman, 1970).

There have been times, such as the 1920's, when the focus of attention was on identifying that knowledge of most worth and developing efficient ways of transmitting it. A similar emphasis was primary during the 1970's when major concern focused on identifying minimum essentials and criterion referenced test items to measure whether or not those essentials had been learned. But there have also been other times, such as the 1940's and 1950's, when concern about the interrelationships among fields of knowledge and ability to understand and use knowledge in the solution of meaningful human problems was occupying the attention of those on the frontiers of education.

In the current movement toward emphasis on thinking skills as goals of education, California has been in the forefront (Paul, 1984). Special instruction in problem solving skills is being instituted in colleges in the state and also in some of the elementary and secondary schools. It is true that the major focus of instruction in the schools is still on traditional subject fields, and the current state tests are focused on traditional school subjects -- reading, written expression, math, social studies and science. But a recent requirement is that one-third of the questions in the state examinations at the eighth grade level be designed to test critical thinking (Pandey, 1984).

In the design of such tests, attention will undoubtedly be paid to developments reported by cognitive psychologists. It may well be that insights based on their findings will enable test developers in the 80's

to come up with unique test questions capable of pointing up specific differences between the thought processes of pupils who are successful in solving thought problems and those who aren't. On the other hand, it may be that at the point of application, critical thinking or problem solving ability is more than the sum of a number of general abilities identified in psychological studies and closely dependent on background knowledge in the content area of the problem.

To the extent that this is the case, it may not be a waste of time to examine the efforts of test developers in years past. Some promising approaches that may have been abandoned when the objective was to provide measures on individuals in large scale programs might turn out to be feasible under the matrix sampling model of a California Assessment Program. Or an approach that proved suitable in an individual institution might, with some creative imagination, be adapted to a wider range of ability or a different age level. There may well be some promising suggestions to be found in archives long closed if only one were willing to take the time to look.

It is not surprising that often the stimulus material for test items designed to measure thinking skills have been reading passages. As early as 1917, E. L. Thorndike was calling attention to the fact that learning to read was a complex task requiring reasoning ability. As early as 1929 Speer was developing measures of appreciation in poetry, prose, and art that are still appropriate for today's use. Neither Thorndike nor Speer depended on multiple-choice questions, but until the development of computer programs for analyzing discourse, the analysis of such data was prohibitively

expensive for large scale testing programs. A survey of the literature might turn up other examples of early efforts that might be useful given today's technology.

In the history of testing in education one name stands out, and there is no better place to begin looking for promising ideas than in the work of Ralph Tyler (1930a, 1930b, 1932a, 1932b, 1940) and some of his associates (Sampson & Tiffany, 1932). It is sobering for the test developer of the 80's to study his work and to realize that the test development process Tyler used at Ohio State University in the early 1930's has not been improved upon. In fact, the procedures used today are often poorer because today's test developers often begin with assumptions of what is and what is not possible that never occurred to Tyler. As reports indicate, he always began by attempting to gather direct evidence of the behavior he was trying to assess, and only after he had successfully done so did he make the attempt to move to a more indirect approach that might provide more reliable, more efficient, or more objective measurement. For example, in attempts to measure student behavior involved in scientific method, Tyler (1932a) identified three essential abilities: the ability to formulate a reasonable generalization from specific scientific data, the ability to plan an experiment which would be used to determine whether or not a proposed hypothesis is true, and the ability to apply general principles to concrete situations. He first designed free response questions for each and then attempted to design objective questions that would provide the same kinds of information more efficiently. In most cases the first attempts were unsuccessful. Thus, when examiners listed scientific data

and optional generalizations from which students were to select the best, the results were not at all like those from free response tests. However, when the optional generalizations were selected from those originally written by the students in the free response situation, and when students were asked to identify both the best generalization and the worst, the correlation with free response scores was .85. On the other hand, the attempt to test ability to draw inferences from facts using an objective test format was not successful. Correlations of the free response test with a parallel multiple-choice test was only .38 and with a test of factual knowledge was only .29 (Tyler, 1930a). Sometimes Tyler succeeded and sometimes he didn't. But apparently he never willingly substituted reliability for validity, as many subsequently have. Tyler demonstrated that sometimes it was possible to obtain better evidence in the time available by using simulations or response type items, and the approaches he reported are still appropriate in many cases today.

A second source of ideas can be found in the work of the Board of Examiners at the University of Chicago. The Manual of Examination Methods (University of Chicago, 1937) illustrates both essay and objective type items that were used during the early years, some of which are clearly designed to test more than simple recall. Of particular interest to those proposing to develop free response questions today, after many years of emphasis on objective test questions, is the section devoted to the essay form (pp. 112-132). However, those interested in testing complex thinking skills with objective questions will find many of the other sections of the Manual filled with evidence of how far the evolution of the multiple-choice format has obscured other promising objective test formats.

Christine McGuire (American Council on Education, 1959, p. 141) describes the program in later years, after Tyler had moved from Ohio State to Chicago to head the Board:

> The questions are designed to require the student to apply concepts and principles to new situations ranging from the analysis in the Humanities Comprehensive of an unfamiliar sonata or painting to the interpretation in the Natural Science Comprehensive of a scientific paper or report of an experiment new to the students and to the evaluation in the Social Science Comprehensive of specific governmental policies not discussed in the relevant course (p. 141).

Once the examinations were administered, they were placed in the library where students could examine them to see the kinds of performance that were required. There might be some profit in searching the archives at Chicago, but my own estimate is that a careful study of the 1937 Manual and of illustrative items from the published work of test developers trained at The University of Chicago or at one of the other similar boards of examiners that were operating in the 40's and 50's (Adkins, 1947; Bloom, 1964; Dressel & Associates, 1961; Engelhart, 1942; Michigan State, 1949) would cover any promising ideas to be found from these sources.

The task of creating effective test questions requiring high level thinking has never been expressed better than by Atkins:

> The construction of a good test is not a task to be tossed off in a few hours by "thinking up" items or by excerpting passages from pertinent printed materials. In fact, experienced item constructors consider 5 to 15 acceptable multiple-choice items per day, on ordinary subject matter, to be a good average-<u>after</u> the basic orientation and planning are completed. The number constructed per day will depend in part upon the nature of the items. For vocabulary and computation items, a good average is 15 to 20 acceptable items per day. Only three or four difficult judgment items in a complex field might be considered satisfactory production. By acceptable items at the point of initial construction is meant not items which are ready for use

> in a test but those which are sufficiently promising to warrant review by subject-matter specialists or tryout on a representative group of subjects.
>
> The construction of test items is essentially a highly creative task of selecting content that will be appropriate and likely to yield a discriminating item; of developing that content into a statement of a problem, and perhaps several alternative solutions to the problem; and then of expressing the item in conformance with the technical principles which yield the most accurate predictions of job success. This work requires the bringing together of the three abilities which are essential to the production of good tests, viz, comprehensive knowledge about and understanding of the field of work or occupational area concerned, skill in the application of the principles of personnel measurement, and facility in the use of words to express ideas. Since all of these abilities are very seldom found in one individual, written tests are often constructed by a team of individuals who bring together the requisite background and abilities. Anyone who finds the English language and the idiom of his special field inadequate to express his thoughts had best not attempt to write test items by himself (pp. 45, 46).

Adkins was focusing on the development of objective type test questions, but good free-response questions are no easier to construct.

Another source of ideas for questions evaluating educational outcomes other than knowledge is the work of the Eight Year Study of the Progressive Education Association (Smith, Tyler, & the Evaluation Staff, 1942). The study produced a number of different measures of aspects of thinking and of aspects of appreciation, and these are described in some detail in the report. Copies of the instruments were, for a time, distributed by Education Testing Service; they may still be available in the ETS archives.

The Eight Year Study report did not have the expected impact. In part, this may be due to the fact that the report was issued during World War II when the focus on education effort was on narrower and more immediately apparent objectives. But it is also true that the tests, which were designed

to measure abilities that did not depend on instruction in any particular school subject, could not readily be related to school learning organized around traditional subject matter fields. It is too early to judge whether or not current efforts of these kinds will be any more successful.

An effort more closely related to traditional content areas was that of the Cooperative Study of Evaluation in General Education (Dressel, 1954; Dressel & Mayhew, 1954a, 1954b, 1954c, 1954d). In that study some effort was devoted to pervasive objectives of education that cut across subject matter, critical thinking and attitudes; but the major focus was on the measurement of outcomes in general subject areas -- social science, communication, science, and the humanities. The reports describe and give examples of the kinds of stimulus material and questions developed in each area. Both choice-type and free-response-type questions were developed, and to a considerable extent the focus was on critical thinking (Dressel & Mayhew, 1954b, pp. 272-278). As in the case of tests developed in the Eight Year Study, those developed in the Cooperative Study of Evaluation in General Education were made available for a time through Education Testing Service and probably can be found in the ETS archives. However, since subtle differences in emphasis occur over time, it is the ideas about approaches rather than the specific total products that are of significance today, and most of these can be gleaned by examining the illustrative material in the published volumes.

The influence of Tyler, particularly during the 1930's and 1940's was primarily on measurement and evaluation at the undergraduate college level. A comparable pioneer contribution at the elementary and secondary

school level was being made by E. F. Lindquist (1970) from his base at The University of Iowa. In 1931 he was placed in charge of the Iowa Every-Pupil Achievement Testing Program operated by the University for secondary schools in the state. In describing the program he wrote, "We tried particularly to emphasize in the test what we called the reasoned understanding of broad concepts and large units of comprehension, and the ability to do some critical thinking about the materials studied, and we minimized the testing of detailed, specific information" (p. 9). Over the years the program has evolved, but the focus established by Lindquist at the beginning is still central to the thinking of the staff responsible for developing new forms. The most recent forms of the Iowa Tests of Educational Development (Feldt, Forsyth, & Alnot, 1986), which replaced the Iowa-Every-Pupil Achievement Tests in 1942, provide appropriate models for those seeking examples of questions requiring thought about, rather than recall of, content in broad academic areas.

A comprehensive account of Lindquist's thinking may be found in Chapter 9 of the comprehensive volume he edited in 1951, Educational Measurement. Lindquist's argument cannot be summarized in the space available here. However, it can be noted that he believed that professional teachers should be free to use a variety of content, depending on the characteristics of their students and their own special backgrounds. He held that tests designed for large scale use should focus, not on specific content, but rather on the generalized abilities that remained after the content details encountered during instruction might be forgotten.

One of the problems recognized by those concerned with the measurement of abilities beyond recall of information has been the fact that the extent to which an individual's answer to a question reflects higher order thinking depends on the examinee's previous instruction. Is the question a new one for the individual or is the question one that has been encountered before and the answer recalled from direct instruction? Tyler recognized the problem while still at Ohio State, and Engelhart (1942) discussed the problem in detail. Perhaps the reason such questions are often found in tests developed in particular colleges and less often in tests developed for widespread use is that the test developer in a particular institution could know about the experiences of the students during instruction. Thus, the college examiner could have a reasonable expectation that a question designed to elicit critical thinking would be one the student had not previously encountered nor one based on material that all students had been assigned to study in anticipation of the test but that had not been discussed in detail in a class.

In the case of tests designed for widespread use, such as the California Assessment Program, however, it is difficult to see how such expectations could be held. For such purposes, the practice has usually been to incorporate background material directly into the questions as described by Ebel (Lindquist, 1951, pp. 241-249). This is not always satisfactory. To allow enough time for the individual to thoroughly comprehend the material is usually unfeasible, yet thorough understanding may be essential to any sound critical thinking (Bentley, 1950). Furthermore, unless the material is new for each testing, problems of security may be

unsolvable. This may account for the fact that good models of thinking exercises are more often found in archives than in operational tests. Not only are such exercises time-consuming and expensive to prepare, but continuing financial support for sustained effort over time is difficult, if not impossible, to obtain.

For examples of test questions designed to measure higher order outcomes of instruction in particular subject fields, one can examine a variety of published articles and volumes sponsored by professional associations such as the National Council of Teachers of English (Fagan, 1975), the National Council of Teachers of Mathematics (1961), the National Council for the Social Studies (Berg, 1965; Morse & McCune, 1964), and the National Society for the Study of Education (Henry, 1946). More general surveys have been sponsored by the American Council on Education (Coffman, 1971; Fitzpatrick & Morrison, 1971; Lindquist, 1951), by the American Educational Research Association (Coffman, 1969; Ebel & Damrin, 1960) and by the New Jersey Department of Education (Boyd & Shimberg, 1971). Most of the readily available examples of ideas for test questions can be traced through these sources.

There are, however, some additional developments that warrant special mention. In 1963 the Education Testing Service published a small booklet, Multiple-choice Questions, a Close Look that was designed to dispel the impression that objective tests could measure only recall or simple types of convergent thinking. Other examples are to be found in ETS archives. In the early 1950's a general education battery, the AREA Tests in science, social studies, and the humanities, were developed for use in the Graduate

Record Examinations. Somewhat later in the decade, an experimental test battery, the Tests of Developed Abilities was developed under the sponsorship of the College Entrance Examination Board (Coffman, 1957). During this same period, ETS provided support services for a study at Oberlin College where the tests were designed to measure as directly as possible outcomes of college experiences in music and art. These tests made use of high fidelity recorded music and color slides of selected paintings. They were designed for administration using matrix sampling because of the extreme length of the tests, and were, for a time, distributed by ETS (Oberlin Tests of Music and Art, no date). Two other experimental measures developed by ETS staff members, The Russell Sage Social Relations Test (Damrin, 1959) and the Inbasket Test (Frederiksen, Saunders, & Wand, 1957) are examples of simulations worth examining.

Finally, during the first administration of the tests of the National Assessment of Educational Progress there was a greater variety of types of exercises than in subsequent years when financial considerations led to the use of greater numbers of machine scorable responses. Of particular interest is the Multi-Topic Discussion exercise used in the Citizenship assessment (National Assessment, 1970, pp. 69-73) and free response items in the Social Studies Tests (National Assessment, 1975).

REFERENCES

Adkins, D. (1947). Construction and analysis of achievment tests. Washington: U.S. Government Printing Office.

American Council on Education, Committee on Measurement and Evaluation. (1959). College Testing. Washington: Author.

Anderson, J. R. (1981). Cognitive skills and their application. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bentley, E. T. (1950). Measurement of some aspects of critical thinking. Journal of Experimental Education, 18, 263-278.

Berg, H. D. (Ed.). (1965). Evaluation in social studies. Washington: National Council for the Social Studies.

Bloom, B. S. (Ed.). (1964). Taxonomy of educational objectives. Handbook I: cognitive domain. New York: McKay.

Bloom, B., Hastings, T., & Madaus, G. (1971). Handbook of formative and summative evaluation of student learning. New York: McGraw-Hill.

Boyd, J. L., Jr., & Shimberg, B. (1971). Handbook of performance testing. Princeton, NJ: Educational Testing Service.

Brandt, R. (Ed.). (1984). When teachers tackle thinking skills. Educational Leadership, 42(3), 3-72.

Brandt, R. (Ed.). (1985). Direct instruction and teaching for thinking. Educational Leadership, 42(8), 3-47.

Coffman, W. E. (1957). The Tests of Developed Ability - Developing the tests. College Board Review, 31, 6-10.

Coffman, W. E. (1966). On the validity of essay tests of achievement. Journal of Educational Measurement, 3, 151-156.

Coffman, W. E. (1969). Achievement tests. In R. L. Ebel (Ed.), Encyclopedia of educational research (pp. 7-17). New York: Macmillan.

Coffman, W. E. (1970). Concepts of achievement and proficiency. Proceedings of the 1969 Invitational Conference on Testing Problems (pp. 3-11). Princeton, NJ: Educational Testing Service.

Coffman, W. E. (1971). Essay examinations. In R. L. Ebel (Ed.), Educational measurement (2nd ed.) (pp. 271-302). Washington: American Council on Education.

Damrin, D. E. (1959). The Russell Sage Social Relations Test: a technique for measuring group problem solving skills in elementary school children. Journal of Experimental Education, 28, 85-99.

Dressel, P. L. (Ed.). (1954). Evaluation in general education. Dubuque, IA: W. C. Brown.

Dressel, P. L. & Associates. (1961). Evaluation in higher education. Boston: Houghton Mifflin.

Dressel, P. L. & Mayhew, L. B. (1954a). Critical thinking in social sciences. Dubuque, IA: W. C. Brown.

Dressel, P. L. & Mayhew, L. B. (1954b). General education: explorations in evaluation. Washington, DC: American Council on Education.

Dressel, P. L. & Mayhew, L. B. (1954c). Handbook for theme analysis. Dubuque, IA: W. C. Brown.

Dressel, P. L. & Mayhew, L. B. (1954d). Science reasoning and understanding. Dubuque, IA: W. C. Brown.

Ebel, R. L. & Damrin, D. E. (1960). Tests and examinations. In C. E. Harris (Ed.), Encyclopedia of educational research (3rd ed.)(pp. 1502 -1517). New York: Macmillan.

Educational Testing Service. (1963). Multiple-choice questions: a close look. Princeton, NJ: Author.

Engelhart, M. D. (1942). Unique types of achievement test exercises. Psychometrika, 7, 103-116.

Fagan, W. T. (Ed.). (1975). Measure for research and evaluation in the English language arts. Urbana, IL: ERIC Clearinghouse on Reading and Communication Skills and National Council of Teachers of English.

Feldt, L. S., Forsyth, R. A., & Alnot, S. D. (1986). Iowa Tests of Educational Development, Forms X-8 & Y-8. Iowa City, IA: Iowa Testing Programs, University of Iowa.

Fitzpatrick, R. & Morrison, E. J. (1971). Performance and product evaluation. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.)(pp. 231-270). Washington, DC: American Council on Education.

Frederiksen, N., Saunders, D., & Wand, B. (1957). The Inbasket Test. Psychological Monographs, 71 (9, Serial No. 438).

Furst, E. J. (1958). Constructing evaluation instruments. New York: Longmans Green.

Gerberich, J. R. (1952). Specimen objective test items: A guide to achievement test construction. New York: Longmans.

Godshalk, R., Swineford, F., & Coffman, W. E. (1966). The measurement of writing ability. New York: College Entrance Examination Board.

Hawkes, H. E., Lindquist, E. F., & Mann, C. R. (1936). The construction of achievement examinations. Boston: Houghton Mifflin.

Henry, N. B. (Ed.). (1946). Measurement of understanding (Forty-fifth yearbooks, Part I, National Society for the Study of Education). Chicago: University of Chicago Press.

Judd, C. H. (1936). Education as cultivation of the higher mental processes. New York: Macmillan.

Lindquist, E. F. (1970). The Iowa Testing Programs--A retropsective view. Education, 91(1), 7-23.

Lindquist, E. F. (Ed.). (1951). Educational measurement. Washington, DC: American Council on Education.

Michigan State College of Agricultural and Applied Science, Board of Examiners. (1949). Comprehensive examinations in a program of general education. East Lansing: Michigan State College Press.

Morse, H. T. & McCune, G. H. (1964). Selected items for the testing of study skills and critical thinking. Fourth Edition: Bulletin No. 15. Washington, DC: National Council for the Social Studies.

National Assessment of Educational Progress. (1975, September). Released social studies exercises (1971-72). Denver: Education Commission of the States.

National Assessment of Educational Progress. (1970, November). Report 2, Citizenship: national results. Washington, DC: U.S. Government Printing Office.

National Council of Teachers of Mathematics. (1961). Evaluation in mathematics (Twenty-Sixth Yearbook). Washington, DC: Author.

Oberlin Tests of Music and Art. (No date). Distributed for Oberlin College by Educational Testing Service, Princeton, NJ, for experimental purposes only.

Pandey, Tej N. (1984, April). Structure of problem solving assessment in California Assessment Program's Survey of Academic Skills: Grade 8. Paper presented at the American Educational Research Association meeting, New Orleans.

Paul, Richard W. (1984, September). Critical thinking: fundamental to education in a free society. Educational Leadership, 42(1), 4-14.

Sampson, H. C. & Tiffany, L. H. (1932). The methods followed in the teaching of general biology. In Ralph W. Tyler, Service studies in higher education. Bureau of Educational Research Monographs, Number Fifteen (pp. 3-42). Columbus: Ohio State University.

Scates, D. E. (1947). Fifty years of objective measurement and research in education. Journal of Educational Research, 41, 241-64.

Smith, E. R., Tyler, R. W., & the Evaluation Staff. (1942). Appraising and recording student progress. New York: Harper.

Speer, R. K. (1929). Measurement of appreciation in poetry, prose, and art, and studies in appreciation. (Contribution to Education, No. 362). New York: Bureau of Publications, Teachers College, Columbia University.

Sternberg, R. J. (1985a). Teaching critical thinking, Part 1: Are we making critical mistakes? Phi Delta Kappan, 67,194-198.

Sternberg, R. J. (1985b). Teaching critical thinking, Part 2: Possible solutions. Phi Delta Kappan, 67, 277-280.

Thorndike, E. L. (1917). Reading reasoning: a study of mistakes in paragraph reading. Journal of Educational Psychology, 8, 323-332.

Tyler, R. W. (1930a). Measuring the ability to infer. Educational Research Bulletin, 9, 475-480.

Tyler, R. W. (1930b). A test of skill in using a microscope. Educational Research Bulletin, 9, 493-496.

Tyler, R. W. (1932a). Ability to use scientific method. Educational Research Bulletin, 11, 1-9.

Tyler, R. W. (1932b). The construction of examinations in botany and zoology. In Ralph W. Tyler, Service studies in high education. Bureau of Educational Research Monographs, Number Fifteen (pp. 43-51) Columbus: Ohio State University.

Tyler, R. W. (1934). Constructing achievement tests. Columbus: Ohio State University.

Tyler, R. W. & White, S. H. (Chairman). (1979, October). Testing, teaching, and learning Report of a Conference on Research on Testing, August 17-26, 1979). Washington, DC: National Institute of Education, Department of Health, Education and Welfare.

University of Chicago, Board of Examinations. (1937). Manual of examination methods (2nd ed.). Chicago: University of Chicago Book Store.

Whimbey, A. (1984). The key to higher order thinking is precise processing. Educational Leadership, 42(1), 66-70.